Genet. Res., Camb. (2001), **78**, *pp.* 49–57. *With 2 figures.* © 2001 Cambridge University Press
DOI: 10.1017/S0016672301005183 Printed in the United Kingdom

Simulating genealogies of selected alleles in a population of variable size

MONTGOMERY SLATKIN*
Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA

(*Received 24 August 2000 and in revised form 21 February 2001*)

Summary

An importance-sampling method is presented that allows the simulation of the history of a selected allele in a population of variable size. A sample path describing the number of copies of an allele that arose as a single mutant is generated by simulating backwards from the current frequency until the allele is lost. The mathematical expectation of a quantity or statistic is then estimated by taking averages over replicate simulations, weighting each replicate by the ratio of its probabilities under the Markov chains for the forward and backwards processes. This method was used to find the average age of a selected allele in an exponentially growing population. In terms of the effect on average allele age, selection in favour of an allele is not equivalent to exponential growth. To generate gene genealogies of a sample of copies of a selected allele, the neutral coalescent model is simulated for the subpopulation containing only the selected allele. From the resulting intra-allelic genealogy, it is possible to calculate the likelihood of the selection intensity as a function of the observed level of variability at marker loci closely linked to the selected allele. This method was used to estimate the intensity of selection affecting the Δ32 allele at the CCR5 locus in Europeans and a mutant at the MLH1 locus associated with colorectal cancer in the Finnish population.

1. Introduction

Coalescent theory, introduced by Kingman (1982), and the nearly equivalent theory of lines of descent introduced by Griffiths (1980), have provided powerful ways to analyse the rapidly increasing body of genetic data from human and other populations. The essential feature of coalescent theory is that, for neutral alleles, the ancestry of only the sample of genes need be modelled. Coalescent theory provides a way to simulate efficiently the history of a sample of neutral alleles in large populations and to estimate likelihoods of population genetic parameters. Griffiths & Tavaré (1994*a*, *b*), Kuhner *et al.* (1995, 1998) and others have developed coalescent-based programs that estimate mutation rates, recombination rates and population growth rates from samples of neutral alleles.

Natural selection has been difficult to incorporate into coalescent theory because the essential simplicity of the neutral coalescent is lost. The gene genealogy of a sample of selected alleles depends not only on the ancestors of copies sampled but also on the history of the whole population. Several approaches have been taken to overcoming this difficulty and carrying out a coalescent analysis of selected alleles. Hudson & Kaplan (1988) modelled the effect of overdominant selection by assuming that it is so strong that the frequencies of the two alleles are fixed – an assumption that reduces the problem to a two-island model with mutation and recombination playing the role of migration. Kaplan *et al.* (1989) used a similar approximation to model genetic hitchhiking. They assumed that the substitution of an advantageous allele is described by deterministic theory and then allowed for mutation and recombination to modify sites linked to advantageous alleles.

Neuhauser & Krone (1997) and Krone & Neuhauser (1997) developed a general method for simulating the genealogy of a sample of alleles subject to selection and mutation. Their method generates a large network, called the ancestral selection graph, in which the genealogy is embedded. As originally presented, their method was impractical for even moderately strong selection because the ancestral selection graph became too large. Slade (2000) has improved the Krone–Neuhauser method in a way that allows for

* Fax: +1 (510) 643 6264. e-mail: slatkin@socrates.berkeley.edu

stronger selection. At present, the approach taken by Krone, Neuhauser and Slade assumes a population of constant size.

J. Felsenstein, M. K. Kuhner and J. Yamato (personal communication) are developing a different method for incorporating selection into a coalescent model in a population of variable size. They use a Markov Chain Monte Carlo (MCMC) method to sample from the distribution of past allele frequency curves for the selected locus, and also to sample from the coalescent of copies of the locus conditional on that selection curve. Both the selection curve and the coalescent of copies are updated by MCMC sampling.

In this paper I will introduce another method, one that relies on importance sampling, for simulating the genealogies of selected alleles. This method differs from that of Krone and Neuhauser and of Felsenstein, Kuhner and Yamato in several ways. It assumes only two alleles and ignores mutation, other than the unique event that created the mutant allele. It simulates the history of the mutant allele from the time it arose until it is found in a specified frequency in the population or in a sample from the population. It allows for arbitrary selection and for arbitrary changes in past population size. The method complements the method of Slatkin & Bertorelle (2001), which tests for selection based on the extent of variability among different copies of an allele at linked marker loci and which estimates population growth rate under the assumption of neutrality.

The importance-sampling method works well for selection of any strength in favour of an allele and for weak selection against it, but performs poorly when strong negative selection is assumed. To examine negative selection affecting low-frequency alleles, the method described by Wiuf (2001) based on a linear birth–death process is preferable. It is unlikely that a method will be needed to analyse strongly deleterious alleles found in high frequency because such alleles would be very unlikely to be present.

2. Importance sampling

(i) *The forward process*

The model assumes a diploid species for which the population size in generation t is N_t. In a particular generation, T, the whole population is surveyed and i copies of an allele A are found. The locus is biallelic, with other allele being a, and the relative fitnesses of the three genotypes are $1+s_1$ (AA), $1+s_2$ (Aa) and 1 (aa), where s_1 and s_2 can take any values greater than -1. All copies of A are descended from a mutation that occurred in generation 0. The sample path describing the numbers of copies of A is denoted by $H = \{i_0, i_1, i_2, \ldots, i_{T-1}, i_T\}$, where $i_0 = 0$, $i_1 = 1$ and $i_T = i$.

The probability of a sample path is found by assuming a Markov chain for transitions from one generation to the next and then multiplying the transition probabilities. Here I assume a Wright–Fisher model with selection, for which the distribution of the number of copies of A is binomial:

$$\Pr(i_t \mid i_{t-1}) = p_{i_{t-1}, i_t} = \binom{2N_t}{i_t} x_{t-1}^{'i_t}(1-x_{t-1}^{'})^{2N_t - i_t}, \tag{1a}$$

where

$$x_{t-1}^{'} = x_{t-1}\frac{1+s_1x_{t-1}+s_2(1-x_{t-1})}{1+s_1x_{t-1}^2+2s_2x_{t-1}(1-x_{t-1})} \tag{1b}$$

and $x_{t-1} = i_{t-1}/(2N_{t-1})$. That is, x_{t-1} is the frequency of A in generation $t-1$ before selection and $x_{t-1}^{'}$ is the frequency in the (infinite) gamete pool after selection. The probability of a sample path H is then

$$\Pr_{\mathrm{F}}(H) = \prod_{t=1}^{T} p_{i_{t-1}, i_t}, \tag{2}$$

where $p_{0, i_1} = 1$ if $i_1 = 1$ and 0 otherwise. The subscript F indicates that this is the probability of the sample path generated by the forward process, which is the process beginning with a newly arisen copy of A at $t = 1$ and continuing until $t = T$.

To study the history of A, we would like to draw randomly from the set of sample paths satisfying our conditions that $i_0 = 0$, $i_1 = 1$ and $i_T = i$. One way to do that is with a rejection scheme: begin with one copy at $t = 1$, simulate forward in time using the Wright–Fisher model until $t = T$, and then reject all sample paths for which $i_T \neq i$. The rejection method works in principle but is impractical for all but very small population sizes. For larger populations, the probability that $i_T = i$ would be so small in general that almost all sample paths would be rejected.

An alternative to a rejection method is importance sampling. A sample path H can be generated by a model that assumes i copies of A at $t = T$ and proceeds backwards in time until A is lost. This sample path can be regarded as a sample path of the forward process by counting time backwards from T and defining $t = 0$ to be the first generation in which no copies of A remain. I will call this way of generating a sample path the backwards process and define $\Pr_{\mathrm{B}}(H)$ to be the probability of a sample path H under the backwards process, which has not yet been specified.

Once a sample path H is obtained it can be used to calculate some quantity or statistic, say $G(H)$, which could be allele age or the probability of observing a configuration of alleles at linked marker loci. If sample paths could be drawn randomly from the forward process, the approximate average of $G(H)$

would be found by randomly choosing a large number of sample paths, computing G for each one, and averaging the result:

$$E(G) = \frac{1}{M} \sum_{m=1}^{M} G(H_m),$$

where M is the number replicates and H_m is the mth sample path. If the backwards process is used to generate sample paths, the contribution of each sample path must be weighted by the ratio of probabilities of H under the forward and backwards processes and by the relative probability that a mutation occurred T generations in the past:

$$E(G) = \frac{\sum_m w_m G(H_m)}{\sum_m w_m}, \tag{3}$$

where

$$w_m = \frac{\Pr_F(H_m)}{\Pr_B(H_m)} N_0 \tag{4}$$

and N_0 is the population size in the first generation in the backwards process after the allele is lost. The term N_0 is needed because the rate of influx of new mutations is proportional to the population size (Slatkin & Rannala, 1997; Wiuf, 2001). If the mutation rate does not vary with time, its value will cancel from all calculations and hence is not needed.

Equation (3) is an example of importance sampling (Tanner, 1993). It will provide an accurate approximation to $E(G)$ if it is possible to generate sample paths so that those with relatively high probabilities under the forward process are generated frequently enough by the backwards process. An indication of how well the method performs is the ratio of the sum of the weights to the maximum weight, $w_{\max}$:

$$W = \sum_{m=1}^{M} \frac{w_m}{w_{max}}, \tag{5}$$

which is necessarily between 1 and M. It takes the maximum value only if all the weights are equal and is 1 if all the weight is given to one of the M replicates. A large value of W does not guarantee that an accurate estimate of $E(G)$ is obtained, but it does indicate that numerous replicates are making a significant contribution to the expected value. A small value of W indicates that the method is not performing well and that the result is probably not accurate because only a few sample paths contribute significantly to the average.

(ii) *Reversibility in a population of constant size*

At this point it is useful to consider the reversibility of Markov chains. A Markov chain is reversible if, for any sample path, $\Pr(i_t, i_{t+1},\ldots, i_{t+n}) = \Pr(i_t, i_{t-1},\ldots, i_{t-n})$ for all t and n. If p_{ij} is the transition matrix and α_i is the stationary distribution, the Markov chain is reversible if and only if it has a stationary distribution and $\alpha_i p_{ij} = \alpha_j p_{ji}$ for all i and j (Ewens, 1979, p. 74). The Moran model with selection and mutation is reversible, and Watterson (1976) used that fact to show that in a Moran model the probability distribution of times to loss of an allele is the same as the distribution of allele age. The Wright–Fisher model in a population of constant size is not reversible, but it is in the diffusion limit because it leads to the same diffusion equation as does the Moran model (Watterson, 1977). Hence, in the diffusion limit, the distribution of allele ages in a Wright–Fisher model is the same as the distribution of times to loss of an allele. The analytical results obtained by Kimura & Ohta (1973), Maruyama (1974) and Li (1975) for average allele age all rely on the reversibility of the diffusion approximation to the Wright–Fisher model.

For a population of constant size, the asymptotic reversibility of the Wright–Fisher model suggests that the Wright–Fisher for the forward process will, when run backwards, generate sample paths that have relatively high probabilities under the forward process. That procedure works for deleterious alleles because in the backwards process they are almost certain to be lost. For advantageous alleles, that approach will not work because such alleles will nearly always be fixed by selection instead of lost, so suitable sample paths representing the increase from a single copy would almost never be obtained. For alleles with an additive effect on fitness, Maruyama's (1974) result provides a solution. Maruyama showed that, in the diffusion limit, the distribution of allele age is invariant to a change in the sign of the selection coefficient. Therefore, for advantageous alleles of additive effect the appropriate backwards process is a Wright–Fisher model with the negative of the selection coefficients. It is necessary to reject all replicates in which $i_1 > 1$, but in practice fewer than half the replicates need be rejected for that reason.

(iii) *Non-additive alleles and variable population size*

For alleles that do not have an additive effect on fitness, the forward process is not invariant to changes in the sign of the selection coefficient. In a population of variable size, the forward process cannot be reversible because the transition matrix of the Markov chain is time-dependent, implying that there is no stationary distribution. In both cases, the choice of the backwards process is not obvious. Equation (3) is true for any backwards process, but it is of practical value only if the backwards process generates sample paths for which the forward probabilities are reasonably large. Otherwise sample paths contributing most

to the expectation will be missed or poorly represented, and the resulting estimate of $G(H)$ will be incorrect. A variety of backwards processes can be envisioned. If selection is weak and population sizes do not vary by much, the choice of the backwards process is not critical, but for strong selection and substantial variation in population size, accurate results are obtained only if the backwards process is chosen carefully.

A useful, but not necessarily optimal, backwards process is a Wright–Fisher model in which the population sizes are in the reverse order and for which the expected allele frequency in the preceding generation is chosen to match the expected change in allele frequency in the forward process as closely as possible. In the backwards process, the sequence of population sizes beginning in generation T is $N_T, N_{T-1}, N_{T-2},\ldots$, and the transition from t to $t-1$ is given by the binomial distribution

$$\Pr(i_{t-1} \mid i_t) = q_{i_t,\, i_{t-1}} = \binom{2N_{t-1}}{i_{t-1}} y_t'^{\,i_{t-1}}(1-y_t')^{2N_{t-1}-i_{t-1}}. \quad (6)$$

For an advantageous allele ($s_1 \geqslant s_2 > 0$), y_t' is chosen so that if $x_{t-1} = y_t'$, then $x_{t-1}' = y_t$ where $y_t = i_t/(2N_t)$. That condition implies $y_t' = z$, where z is the solution to the quadratic equation

$$z\,\frac{1+s_1 z+s_2(1-z)}{1+s_1 z^2+2s_2 z(1-z)} = y_t \quad (7)$$

that lies in the interval $(0, y_t)$. The same choice for y_t' serves for an overdominant allele ($s_2 > s_1$), provided that there is a solution to (7) with z in $(0, y_t)$. For neutral alleles, $y_t' = y_t$.

For a deleterious allele ($s_1 \leqslant s_2 < 0$) the choice of the selection coefficients used for the backwards process is more problematic and there seems to be no choice that works in all cases. For weak selection against A, I found that using z defined by (7) but with $-s_1$ and $-s_2$ replacing s_1 and s_2 provides adequate results if A is in low frequency.

The probability of the sample path H under the backwards process is

$$\Pr_{\mathrm{B}}(H) = \prod_{t=1}^{T} q_{i_t,\, i_{t-1}}. \quad (8)$$

To compute the weighting factor, we take the ratio of the forward and backward probabilities. The result simplifies because almost all the binomial coefficients cancel and most of the remaining terms group conveniently:

$$w = N_0\,\frac{p_{0,i_1}p_{i_{T-1},i_T}}{q_{i_1,0}q_{i_2,i_1}}\prod_{t=2}^{T-1}\left(\frac{x_{t-1}'}{y_{t+1}'}\right)^{i_t}\left(\frac{1-x_{t-1}'}{1-y_{t+1}'}\right)^{2N_t-i_t}. \quad (9)$$

As mentioned above, in the backwards simulation i_1 is not necessarily 1. If it is not, $w = 0$ for that replicate.

3. Average allele age

The importance-sampling method can be tested by finding the average allele age, for which some analytical results are known. The age of an allele is the time since it arose by mutation. To estimate the average age, the backwards process is simulated for each replicate until A is lost or fixed. Replicates in which A is fixed and those in which there are two or more copies in the generation before A is lost are rejected. In the remaining replicates, the time to loss, T_m, and the weight, w_m, in the mth replicate are recorded. After a large number of replicates, the average age is the weighted sum

$$\bar{T} = \frac{\sum_{m=1}^{M} w_m T_m}{\sum_{m=1}^{M} w_m} \quad (10)$$

and a histogram of ages is obtained by binning the ages weighted by w_m. The results presented in Table 1 are based on weighted averages over 100 000 replicates for which non-zero weights were obtained.

Average ages can be compared with expected values obtained from analytical theory for the case of a population of constant size and for neutral alleles in a population of variable size. Kimura & Ohta (1973) showed that the expected age of a neutral allele found in frequency p found in a population of constant size N is approximately

$$E(T) = \frac{-4Np\ln(p)}{1-p}. \quad (11)$$

For selected alleles, diffusion theory provides the solution in the form of an integral that must be evaluated numerically. Maruyama (1974), Li (1975) and Watterson (1976, 1977) have published results for various cases. The expected values in Table 1 are from table 3 of Maruyama (1974).

For a population of variable size, Griffiths & Tavaré (1998) developed a method for finding the expected age and the distribution of ages of a neutral allele. Their analytical expression for expected age is difficult to evaluate numerically but they described an efficient Monte Carlo method which was implemented by Slatkin (2000).

Table 1 presents average allele ages calculated using the importance-sampling method (in roman type), and, when possible, from the relevant analytical theory (in italics). When comparisons can be made, the simulation results are quite accurate. In all cases, W was at least 100 and, for $s \geqslant 0$, at least 1000. The importance-sampling method did not produce usable

Table 1. *Average age (in units of $4N_0$ generations) of an allele in frequency p*

	$4N_0s$					
$4N_0r$		-100	-10	0	10	100
$p = 0{\cdot}01$	0	*0·011*	*0·029*	*0·046*	*0·029*	*0·011*
		0·012	0·030	0·045	0·028	0·012
	10			*0·019*		
		0·011	0·019	0·018	0·017	0·011
	100			*0·0077*		
		0·0082	0·0075	0·0074	0·0073	0·0063
$p = 0{\cdot}1$	0	*0·030*	*0·129*	*0·255*	*0·129*	*0·030*
		0·031	0·129	0·252	0·129	0·031
	10			*0·073*		
		0·032	0·080	0·073	0·064	0·028
	100			*0·021*		
		0·027	0·022	0·021	0·021	0·016
$p = 0{\cdot}5$	0			*0·693*	*0·316*	*0·052*
				0·697	0·317	0·052
	10			*0·160*		
				0·162	0·138	0·047
	100			*0·036*		
				0·036	0·035	0·026
$p = 0{\cdot}9$	0			*0·948*	*0·502*	*0·074*
				0·987	0·501	0·075
	10			*0·203*		
				0·203	0·187	0·067
	100			*0·042*		
				0·042	0·041	0·035

In all cases, $N(t) = N_0 e^{-rt}$ with $N_0 = 10^4$, and the fitnesses $1+2s$ (AA), $1+s$ (Aa), 1 (aa). Results from the importance-sampling method are in roman type; results from analytical theory are in italics. The method did not yield accurate results for $p = 0{\cdot}5$ and 0·9 with negative selection.

results for $s < 0$ for $p = 0{\cdot}5$ or $p = 0{\cdot}9$. Values of W were between 1 and 10 for these cases.

These results show that additive selection and exponential population growth are not equivalent in their effects on average allele age. The reason is that, as pointed out by Wiuf (2000, 2001), although selection and exponential population growth have roughly the same effects on the changes in the numbers of copies of the mutant each generation, population growth but not selection results in an increasing influx of mutations. Table 1 also shows that under strong population growth, the average allele age is not invariant to a change in the sign of the selection coefficient.

4. The gene genealogy of selected alleles

The importance-sampling method assumes that the entire population is sampled, but it can be modified to allow for sampling and can generate the gene genealogy of selected alleles. There are two situations that have to be treated separately. In the first, a random sample of the population is drawn and the frequency of A is known only from that sample. In the second, the frequency of A is estimated from other studies. The second situation is the one usually encountered when studying disease-associated alleles in human populations. The allele frequency is estimated from epidemiological surveys that determine disease prevalence. Then a small sample of chromosomes carrying a particular allele is chosen for more detailed genetic analysis and genotypes are assessed at one or more closely linked marker loci. Often those marker loci played a role in the mapping and cloning of the disease-associated allele and are used subsequently to estimate the age or number of independent origins of the allele (Slatkin & Rannala, 2000). The only difference between these two cases is the way in which the weighting factor is computed. When the allele frequency is estimated from the sample, an additional term is needed.

(i) *Allele frequency estimated from the sample*

Of the $2N_T$ copies of the locus, n are sampled. Consider first the forward process described in Section 2. The allele A arises by mutation in generation 1 and increases to i_T copies in generation T. Then a sample

of size n is drawn without replacement. The distribution of j, the number of copies of A in the sample, is hypergeometric:

$$\Pr_{\mathrm{F}}(j\,|\,i_T) = \frac{\binom{i_T}{j}\binom{2N_T - i_T}{n-j}}{\binom{2N_T}{n}}. \quad (12)$$

In terms of the sample path, H, we can simply add j to the end and say that our sample path is now $\{i_0, i_1, \ldots, i_T, j\}$. The forward probability of this sample path is found by multiplying the product of the transition matrices in (2) by the hypergeometric distribution in (12). To find the probability of this sample path for the backwards process, we need to add a 'reverse sampling' step. That is, we need the probability distribution of i_T given j. A convenient distribution is obtained by assuming the population composition is obtained from a Polya urn model. The process is to draw one gene from the sample and then add a second of the type drawn. The resulting distribution has the same algebraic form as a hypergeometric but is different because the random variable is i_T not j:

$$\Pr_{\mathrm{B}}(i_T\,|\,j) = \frac{\binom{i_T - 1}{j-1}\binom{2N_T - i_T - 1}{n-j-1}}{\binom{2N_T - 1}{n-1}}. \quad (13)$$

This distribution arises naturally in finding the probabilities of descendent configurations in the neutral coalescent (Griffiths & Tavaré, 1998), but here it is used to represent reverse sampling. Other distributions could be chosen instead. In most applications of this method $2N_T \gg n$, in which case the distribution of $x_T = i_T/(2N_T)$, the population frequency of A, is approximately beta:

$$\Pr(x_T\,|\,j) = (n-j)\binom{n-1}{j-1} x_T^{j-1}(1-x_T)^{n-j-1}. \quad (14)$$

To find the appropriate weighting factor for each sample path, we multiply the product in (7) by the ratio of (12) to (13). Most terms cancel leaving

$$\frac{\Pr_{\mathrm{F}}(j\,|\,i_T)}{\Pr_{\mathrm{B}}(i_T\,|\,j)} = \frac{i_T(2N_T - i_T)n}{2N_T j(n-j)}. \quad (15)$$

To simplify further, we note that n, j and N_T are the same for every replicate and so can be ignored. The appropriate weight for each replicate is then obtained by multiplying w in (7) by $i_T(2N_T - i_T)$. If we assume sampling with replacement, implying a binomial distribution of j, and use the beta distribution to approximate (13), the resulting ratio is $x_T(1-x_T)$, which is equivalent because $2N_T$ is the same in every replicate.

(ii) *Allele frequency estimated from other studies*

If the frequency of A is known, then we can proceed by setting x_T to that frequency. No additional factor is needed. The backwards simulation provides a sample path and (9) provides the weight attached to that sample path. Uncertainty in x_T could be incorporated by assuming a distribution of values and then choosing randomly from that distribution at the beginning of each replicate.

(iii) *Allelic genealogy of the sample*

Once a sample path is obtained, we can use the neutral coalescent for a sample of size j in a 'population' of A-bearing chromosomes of sizes $i_T, i_{T-1}, \ldots, i_1$, going backwards in time to generate the gene genealogy of the j copies of A in the sample. The probability of a coalescent event per pair of copies in generation t in the past is $1/(2i_t(i_t-1))$. The result of simulating the neutral coalescent is a gene genealogy for that replicate which represents the genealogical history of the sample of j A-bearing chromosomes (called the intra-allelic genealogy). For many purposes, the $j-1$ coalescence times (called the intra-allelic coalescence times and denoted by $t_2, \ldots, t_j$) of the intra-allelic genealogy are needed. The coalescence time, t_k, is the time at which the number of ancestral lineages in the intra-allelic genealogy increases from $k-1$ to k. When the intra-allelic coalescence times are used to compute a statistic, the expected value of that statistic is obtained by taking the average across replicates weighted by w_m. In each replicate, a new neutral genealogy is simulated using the sample path for that replicate.

5. Applications

I will discuss two applications of this theory. In each, the allele frequency was estimated in other studies, and the problem is to use the allele frequency and the extent of intra-allelic variability to estimate the selection intensity affecting an allele. Both population growth and selection tend to reduce allele age and hence reduce all the intra-allelic coalescence times. When those coalescence times are smaller there is less time during which mutation and recombination can create variation at marker loci linked to the allele of interest. The theory described in the preceding sections provides a way to generate intra-allelic coalescence times for assumed population growth rates and selection intensities. Those coalescence times can be combined with models of mutation and/or recombination to find the probability of the observed level of intra-allelic variability under the model. The two examples illustrate the analysis of two kinds of intra-allelic variability. In the first, the extent of linkage disequilibrium at a closely linked marker locus is

known. In the second, the recombination length of a conserved haplotype is known.

(i) *The probability of the number of non-recombinant chromosomes*

One kind of data that can be analysed with this method is the number of non-recombinant chromosomes in a sample. Typically, a marker locus with two alleles, M and m, closely linked to the allele of interest is surveyed on j chromosomes and a number of them are found to carry one of the alleles, say M, that is in lower frequency on non-A bearing chromosomes. This linkage disequilibrium between A and M provides the basis for linkage disequilibrium mapping when the map location of A is unknown, and for estimating the age of A when the recombination rate between the two loci is known (Slatkin & Rannala, 2000). The data are j, the number of A-bearing chromosomes sampled, and l, the number of AM chromosomes in the sample ($l \leqslant j$). Given the intra-allelic genealogy, the relevant parameters are the per generation rates of M to m transitions and of m to M transitions on A-bearing chromosomes. If the mutation rate from M to m is μ, the rate of mutation from m to M is ν, the frequency of M on non-A-bearing chromosomes is Q, and the recombination rate is c, then the probability that an AM lineage becomes an Am lineage is $u = c(1-Q)+\mu$ and the probability that an Am lineage becomes an AM lineage is $v = cQ+\nu$. The problem is to find the probability of the data, given u, ν and the intra-allelic coalescence times. This probability can be found by simulation for arbitrarily large sample sizes using the method described by Rannala & Slatkin (1998). For relatively small j, a matrix method described by Slatkin (2000) and Slatkin & Bertorelle (2001) is adequate.

The matrix method relies on the independence of changes on each of the k lineages present between t_k and t_{k+1}, where it is convenient to define t_{j+1} to be 0. When there are k lineages present, the state of the system is described by a $k+1$ vector, $p^{(k)}$, for which the lth element is the probability that there are l AM chromosomes present ($0 \leqslant l \leqslant k$). By assuming independence of the changes on each lineage, a $k+1$ by $k+1$ transition matrix, $\mathbf{T}^{(k)}$, can be found such that $\mathbf{T}^{(k)}p^{(k)}$ is the configuration before t_{k+1}, given that $p^{(k)}$ is the configuration immediately after t_k. A second matrix, $\mathbf{S}^{(k)}$, is needed to account for the splitting of one randomly chosen lineage at t_{k+1}. This matrix is a $k+2$ by $k+1$ matrix. $\mathbf{S}^{(k)}$ has as the jlth element $1-j/k$ if $l = j$, j/k, if $l = j+1$, and 0 otherwise. The configuration immediately after t_{k+1}, $p^{(k+1)}$ is $\mathbf{S}^{(k)}p^{(k)}$. By successively multiplying by these matrices and assuming that at t_2, $p^{(2)} = (0, 0, 1)$, meaning that at t_2 both lineages are AM, the probability of l AM chromosomes is found. If s is the unknown parameter, then each intra-allelic genealogy provides $L(s)$, the likelihood. The weighted average of these likelihoods is an estimate of the overall likelihood of s.

(ii) Δ32 *allele of CCR5 in Europeans*

The Δ32 allele of the CCR5 locus is found at a frequency of 10% or greater in all European populations that have been studied (Stephens *et al.*, 1998). It is absent or nearly so from other groups. Individuals homozygous for Δ32 are resistant to infection by human immunodeficiency virus (HIV) and heterozygous carriers who have been infected by HIV have a significantly delayed time to onset of AIDS (Dean *et al.*, 1996). Stephens *et al.* (1998) examined two microsatellite markers closely linked to CCR5 and found most carried a haplotype that is otherwise rare in the population. Based on the extent of linkage disequilibrium with these two marker loci, they concluded that Δ32 is very young, less than 1000 years old, and has been subject to strong positive selection. To estimate the selection coefficient in favour of Δ32, they assumed that it arose by mutation 900 years ago and then fitted a deterministic model of selection. They concluded that, assuming no dominance in fitness, the selection coefficient in favour of Δ32 was roughly 0·3.

We can reanalyse the data of Stephens *et al.* (1998) using the importance-sampling method. Assume additive selection ($s_1 = 2s$ and $s_2 = s$), where s is the parameter to be estimated. A reasonable model of the demographic history of the European population is exponential growth with a current effective size of $N_T = 10^8$ and $r = 0{\cdot}02$ representing the relatively rapid growth in the recent past. Of course, simple exponential growth is not realistic, but if s is an order of magnitude larger than r, the resulting estimate of s will not be very sensitive to the details of the demographic model. Stephens *et al.* (1998) examined 46 chromosomes carrying Δ32 and found at the marker locus denoted GAAT that 44 of them carried the 197 allele, which is at a frequency 0·685 on chromosomes not carrying Δ32. The recombination rate between CCR5 and GAAT is roughly 0·0021. In the notation of the preceding sections, $j = 46$, $l = 44$, $c = 0{\cdot}0021$ and $Q = 0{\cdot}585$. In the likelihood analysis, I followed Stephens *et al.* (1998) in assuming a mutation rate $\mu = 0{\cdot}001$ away from 197 and no mutations creating that allele, meaning that $\nu = 0$. These assumptions imply $u = 0{\cdot}00166$ and $v = 0{\cdot}00144$. The population frequency, x_T, was assumed to be 0·1.

Fig. 1 shows the log of the likelihood of s given these parameter values. We can see that there is a maximum at roughly 0·2, which is close to the value estimated by Stephens *et al.* (1998), but that the decrease of the likelihood with larger s is so slow that stronger selection cannot be excluded.

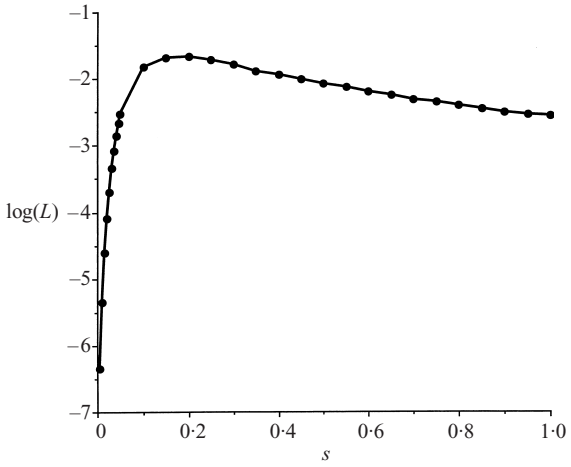

Fig. 1. Estimates of the log-likelihood, $\ln(L)$, of s, the selection intensity, as a function of the data at a marker locus closely linked to the Δ32 allele of CCR5. Values at each point are based on 100000 replicates for which non-zero weights were obtained.

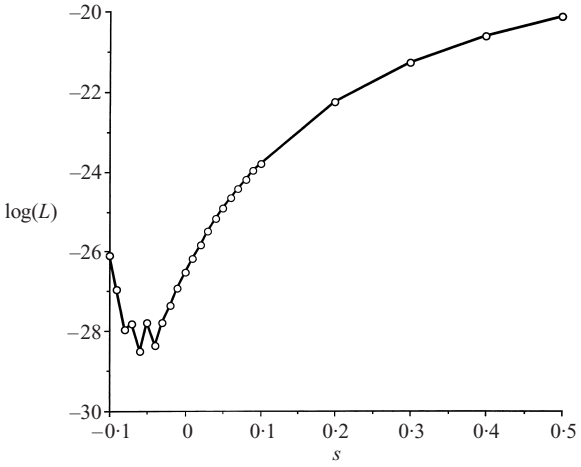

Fig. 2. Estimates of the log-likelihood, $\ln(L)$, of s, the selection intensity, given that a conserved haplotype spanning 7·1 cM surrounds MLH1-1 on all 19 chromosomes examined. As in Fig. 1, values at each point are based on 100000 replicates for which non-zero weights were obtained.

(iii) *Length of conserved haplotype*

Another way to characterize intra-allelic variability is the length of a multilocus haplotype found on all A-bearing chromosomes. The data then are the length of the shared haplotype, l_o, defined to be the number of bases separating the markers at the ends. Slatkin & Bertorelle (2001) show that, given the total length of the intra-allelic genealogy,

$$\Lambda = t_2 + \sum_{k=2}^{i} t_k, \tag{16}$$

the probability of the data is

$$\Pr(l_o) = (\rho\Lambda)^2 l_o e^{-l_o\rho\Lambda}, \tag{17}$$

where ρ is the recombination rate between adjacent bases. Only the product ρl_o, which is the recombination distance separating the ends of the conserved haplotype, affects the results. To find the overall probability of l_o, the weighted average is taken over many replicate sets of intra-allelic coalescence times.

(iv) *Application to MLH1 in Finland*

MLH1 is a DNA mismatch repair gene associated with the elevated incidence of hereditary nonpolyposis colorectal cancer (HNPCC). Moisio *et al.* (1996) sampled 19 chromosomes carrying an allele designated MLH1-1 from the Finnish population and surveyed several closely linked microsatellite loci. They found that a conserved haplotype spanning 7·1 cM was present on all 19 chromosomes. For the calculations, it is reasonable to assume that $\rho = 10^{-8}$ and $l_o = 7{\cdot}1 \times 10^6$.

For the demographic model, I assumed that the current effective population size of Finland is 2000000, that it was founded by a population of effective size 1000, 2000 years or roughly 100 generations ago, and that in 1700 its effective size was about 300000 (S. K. Service & N. B. Freimer, unpublished data). These data can be fitted with a two-stage model of exponential growth, at a rate $r_1 = 0{\cdot}19$ between the present and 1700 (15 generations in the past) and a growth rate $r_2 = 0{\cdot}067$ before 1700. The frequency of MLH1-1 in the Finnish population is difficult to estimate because it is so rare. Moisio *et al.* (1996) estimated the frequency to be 0·0004, but P. Peltomaki (personal communication) found no copies of MLH1-1 in 2351 healthy anonymous blood donors in a region of Finland with the highest frequency of MLH1-1, suggesting a frequency much less than 0·0004. I assumed a frequency x_T of 0·0001 for these calculations.

Fig. 2 shows the log-likelihood of s as a function of selection intensity, assuming $s_1 = 2s_2 = 2s$. These results illustrate the limitation of the importance-sampling method. For $s \geqslant -0{\cdot}01$, the method performs well: the estimated likelihood curve is smooth and values of W are 1000 or greater. For $s \leqslant -0{\cdot}03$, the likelihood curve is no longer smooth and values of W are less than 10, both of which indicate poor performance. The transition occurs at $s = -0{\cdot}02$, for which $W = 79$.

Fig. 1 implies that MLH1-1 has been strongly selected since it arose by mutation but it does not allow us to distinguish between positive and negative selection . The results presented in Table 1 and those derived by Wiuf (2001) show that negative selection can also shorten the intra-allelic genealogy, so the MLH1-1 data are consistent with strong negative selection as well. At this point, it does not seem possible for population genetic analysis to tell us

whether MLH1-1 is deleterious or advantageous to carriers but the role of MLH1 in DNA repair and the association of MLH1-1 with colorectal cancer suggests that it is deleterious. But the associated cancer is probably not the cause of the selection because it is a late-onset condition that has a relatively small effect on reproductive fitness.

6. Discussion and conclusion

The methods described here for simulating the history of a selected allele in a population of variable size have a variety of applications. One application is finding the expected age and the distribution of ages of selected alleles in a population that varies in size. For alleles that have an additive effect on fitness, past population growth reduces average allele somewhat more than does selection of comparable intensity. Maruyama's (1974) result that the age of an additive allele is independent of the sign is not true in a population undergoing exponential growth.

If information about linked marker loci and demographic history is available, this method allows the estimation of selection intensities affecting an allele. The method complements that described by Slatkin & Bertorelle (2001), which provides a test of neutrality and estimates population growth rate if the allele is neutral.

This research was supported in part by a grant from the US NIH, R01-GM40282. I thank R. Nielsen, R. Thomson and C. Wiuf for helpful discussions of the theoretical aspects of this problem and Dr P. Peltomäki for providing information about the frequency of MLH1-1 in the Finnish population.

References

Dean, M., Carrington, M., Winkler, C., Huttley, G. A., Smith, M. W., Allikmets, R., Goedert, J. J., Buchbinder, S. P., Vittinghoff, E., Gomperts, E., Donfield, S., Vlahov, D., Kaslow, R., Saah, A., Rinaldo, C., Detels, R. & O'Brien, S. J. (1996). Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. *Science* **273**, 1856–1862.

Ewens, W. J. (1979). *Mathematical Population Genetics*. Berlin: Springer.

Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theoretical Population Biology* **17**, 37–50.

Griffiths, R. C. & Tavaré, S. (1994*a*). Ancestral inference in population genetics. *Statistical Science* **9**, 307–319.

Griffiths, R. C. & Tavaré, S. (1994*b*). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131–159.

Griffiths, R. C. & Tavaré, S. (1998). The age of a mutation in a general coalescent tree. *Stochastic Models* **14**, 273–295.

Hudson, R. R. & Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840.

Kaplan, N. L., Hudson, R. R. & Langley, C. H. (1989). The 'hitchhiking effect' revisited. *Genetics* **123**, 887–899.

Kimura, M. & Ohta, T. (1973). The age of a neutral mutant persisting in a finite population. *Genetics* **75**, 199–212.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications* (1982), 233–248.

Krone, S. M. & Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology* **51**, 210–237.

Kuhner, M. K., Yamato, J. & Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* **140**, 1421–1430.

Kuhner, M. K., Yamato, J. & Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**, 429–434.

Li, W.-H. (1975). The first arrival time and mean age of a deleterious mutant gene in a finite population. *American Journal of Human Genetics* **27**, 274–286.

Maruyama, T. (1974). The age of an allele in a finite population. *Genetical Research* **23**, 137–143.

Moisio, A. L., Sistonen, P., Weissenbach, J., de la Chapelle, A. & Peltomäki, P. (1996). Age and origin of two common MLH1 mutations predisposing to hereditary colon cancer. *American Journal of Human Genetics* **59**, 1243–1251.

Neuhauser, C. & Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics* **145**, 519–534.

Rannala, B. & Slatkin, M. (1998). Likelihood analysis of disequilibrium mapping, and related problems. *American Journal of Human Genetics* **62**, 459–473.

Slade, P. F. (2000). Simulation of selected genealogies. *Theoretical Population Biology* **57**, 35–49.

Slatkin, M. (2000). Allele age and a test for selection on rare alleles. *Philosophical Transactions of the Royal Society of London*, **B 355**: 1663–1668.

Slatkin, M. & Bertorelle,, G. (2001). The use of intra-allelic variability for testing neutrality and estimating population growth rate. *Genetics* **158**, 865–874.

Slatkin, M. & Rannala, B. (1997). The sampling distribution of disease-associated alleles. *Genetics* **147**, 1855–1861.

Slatkin, M. & Rannala, B. (2000). Estimating allele age. *Annual Review of Genomics and Human Genetics* **1**, 225–249.

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., Winkler, C., Huttley, G. A., Allikmets, R., Schriml, L., Gerrard, B., Malasky, M., Ramos, M. D., Morlot, S., Tzetis, M., Oddoux, C., di Giovine, F. S., Nasioulas, G., Chandler, D., Aseev, M., Hanson, M., Kalaydjieva, L., Glavac, D., Gasparini, P. & Dean, M. *et al.* (1998). Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *American Journal of Human Genetics* **62**, 1507–1515.

Tanner, M. A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Berlin: Springer.

Watterson, G. A. (1976). Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theoretical Population Biology* **10**, 239–253.

Watterson, G. A. (1977). Reversibility and the age of an allele. II. Two-allele models with selection and mutation. *Theoretical Population Biology* **12**, 179–196.

Wiuf, C. (2000). On the genealogy of a sample of neutral rare alleles. *Theoretical Population Biology* **58**, 61–75.

Wiuf, C. (2001). Rare alleles and selection. *Theoretical Population Biology*, in press.